AN INFORMATION SYSTEM FOR CORPORATE USERS: Wide Area Information Servers

by Brewster Kahle
Thinking Machines Corporation
and
Art Medlar
Scolex Information Systems

REPRINTED FROM
September 1991 *ONLINE*®

AN INFORMATION SYSTEM FOR CORPORATE USERS: Wide Area Information Servers

by Brewster Kahle
Thinking Machines Corporation
and
Art Medlar
Scolex Information Systems

To explore text-based information systems for corporate executives, four companies have jointly developed a prototype that gives flexible access to full-text documents. The four participating companies are Dow Jones & Company, Inc. with its premier business information sources; Thinking Machines Corporation, with its high-end information retrieval engines; Apple Computer, with its user interface expertise; and KPMG Peat Marwick, with its information-hungry user base.

One of the primary objectives of the project is to allow a user to retrieve personal, corporate, and wide area information through one easy-to-use interface. For example, instead of using Lotus Magellean for personal information, Verity's Topic for corporate data, and Mead Data Central's NEXIS for published text, one application can access all three categories of information. The user isn't required to become familiar with several entirely different systems. In addition, since the interface consolidates data from many different sources, they can be manipulated effortlessly, virtually without regard to their origins.

The Wide Area Information Server (WAIS, pronounced "ways") project is an experimental venture seeking to determine whether current technologies can be used to create profitable end-user full-text information systems. Fifteen users have been actively using the system for over three months. They have integrated it into their workday routine in much the same way as they have previously integrated spreadsheets and word processors. This preliminary success has convinced us that a WAIS-like system can be a valuable tool for corporate information retrieval. This paper discusses the design and implementation of the prototype system.

THE NEED FOR A WIDE AREA INFORMATION SERVER (WAIS)

Electronic publishing is the distribution of textual information over electronic networks. It has been emerging as a viable alternative to traditional print publishing as the necessary underlying technologies develop. Among the more essential of these are:

- High resolution display screens
- Reliable, high-speed data communication
- Desktop publishing systems
- Inexpensive data storage media

While these technologies have been developed for uses other than electronic publishing, they are the necessary precursors for full-text retrieval systems.

From the user's point of view, there are several problems to be overcome. First, there must be some way of finding and selecting databases from a potentially unlimited pool. Second, although these databases may be organized in different ways, the user should not need to become familiar with the internal configuration of each one. Finally, there must be some practical way of organizing responses on the user's machine to maintain control over what may become a vast accumulation of data.

In addition, developers are faced with a number of architectural issues. The system must be scalable; that is, it must allow for the future growth of both the complexity and number of clients and servers. It must be secure; each server's data must be protected from corruption, and the privacy of the users must be ensured. Lastly, since an unreliable source is useless in a corporate environment, access must be thoroughly robust.

SYSTEM OVERVIEW

The prototype WAIS system takes advantage of current state-of-the-art technology, and presents solutions to all of the above problems. The system is composed of three separate parts: clients, servers, and the protocol that connects them.

The client is the user interface, the server does the indexing and retrieval of documents, and the protocol is used to transmit the queries and responses. The client and server are isolated from each other by the protocol. Any client that is capable of translating a user's request into the standard protocol can be used in the system. Likewise, any server capable of answering a request encoded in the protocol can be used. In order to promote the development of both clients and servers, the protocol specification is public, as is its initial implementation.

On the client side, questions are formulated as English-language questions. The client application then translates the query into the WAIS protocol, and transmits it over a network to a server. The server receives the transmission, translates the received

packet into its own query language, and searches for documents satisfying the query. The list of relevant documents are then encoded in the protocol, and transmitted back to the client. The client decodes the response, and displays the results. The documents can then be retrieved from the server.

THE DIGITAL RESEARCHER

The traditional information research scenario is familiar to anyone who has ever visited a reference desk at a public or corporate library. The client approaches a librarian with a description of needed information. The librarian might ask a few background questions, and then draws from appropriate sources to provide an initial selection of articles, reports, and references. The client sorts through this selection to find the most pertinent documents. With feedback from these trials, the researcher can refine the materials and even continue to supply the user with a flow of information as it becomes available. Monitoring which articles were useful can help keep the researcher on-track.

The WAIS system is an attempt at automating this interaction: the user states a question in English, and a set of document descriptions come back from selected sources. The user can examine any of the items, be they text, picture, video, sound, or whatever. If the initial response is incomplete or somehow insufficient, the user can refine the question by stating it differently.

In addition, the user may also mark some of the retrieved documents as being "relevant" to the question at hand, and then re-run the search. The server recognizes the marked documents, and attempts to find others which are similar to them. In the present WAIS system, "similar" documents are simply ones which share a large number of common words; however, there is potentially no upper limit on the intelligence of a server in determining what similarity entails. This method of information retrieval is called "relevance feedback." The idea has been around for many years [1], and the first commercial system utilizing it, DowQuest [2], was named *ONLINE* Magazine Product of the Year in November 1989.

USER INTERFACES: ASKING QUESTIONS

Users interact with the WAIS system through the Question interface. The interface may appear different on various implementations: for example, a character display terminal will have a different look than one that is capable of displaying bit-mapped graphics. The key, however, is that the user need only become familiar with one interface, which then provides access to all available information sources.

The key, however, is that the user need only become familiar with one interface, which then provides access to all available information sources.

The WAIS system, in this first incarnation, was designed to be used by accountants and corporate executives who are relatively untrained in search techniques. Consequently, to aid these users who have neither the time nor desire to learn a special purpose query language, the system uses English language queries augmented with relevance feedback. While the system's servers currently do not extract semantic information from the English queries, they do their best to find and rank articles containing the requested words and phrases. Used in conjunction with relevance feedback, this method of searching has proven to be more than adequate for the types of searches and databases typically encountered.

The screen illustrations (shown in Steps 1-4 beginning on page 58) are taken from the initial WAIStation program produced at Thinking Machines for the Apple Macintosh. Several other interfaces are under development at Apple Computer, Dow Jones, and elsewhere.

CONTACTING REMOTE SOURCES OF INFORMATION

From the user's point of view, a server is a source of information. It can be located anywhere that one's workstation has access to: on the local machine, on a network, or on the other side of a modem. The user's workstation keeps track of a variety of information about each server. The public information about a server includes how to contact it, a description of the contents, and the cost. In addition, individual users maintain certain private information about the servers they use. Users need to budget the money they are willing to spend on information from particular servers, they need to know how often and when each server is contacted, and they need to assess the relative usefulness of each server. This information helps guide the workstation in making cost effective decisions in contacting servers (Figure 1 on page 60).

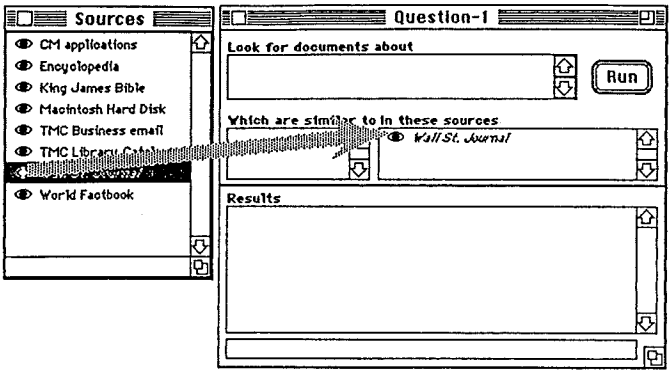
With most current retrieval systems, complications develop as soon as one begins dealing with more than one source of information. The most common problem is that of asking a particular question. For example, one contacts the first source, asks it for information on some topic, contacts the next source, asks it the same questions (most likely using a different query language, a different style of interface, a different system of billing), contacts the next source, and so on. One of the primary motivations behind the initial development of the WAIS system was to replace all this with a single interface.

With WAIS, the user selects a set of sources to query for information, and then formulates a question. When the question is run, the system automatically asks all the servers for the required information with no further interaction necessary by the user. The documents retrieved are sorted and consolidated in a single place, to be easily manipulated by the user. The user has transparent access to a multitude of local and remote databases.

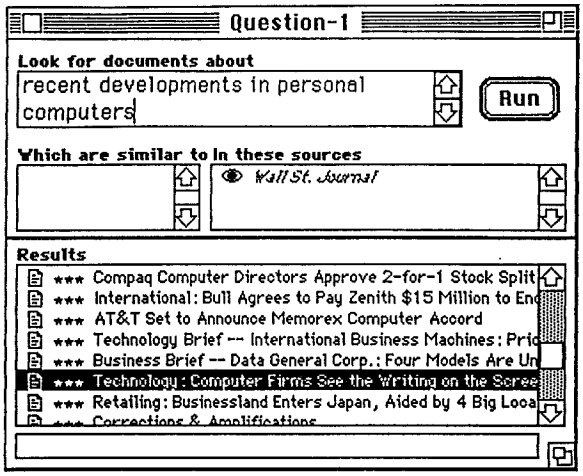
RERUNNING QUESTIONS: A PERSONAL NEWSPAPER

In addition to providing interactive access to a vast quantity of information, the WAIS system can also be used as a rudimentary personal newspaper. A virtually unlimited number of queries can be saved, and updated at periodic intervals. To do this, the user's workstation is directed to contact each server at

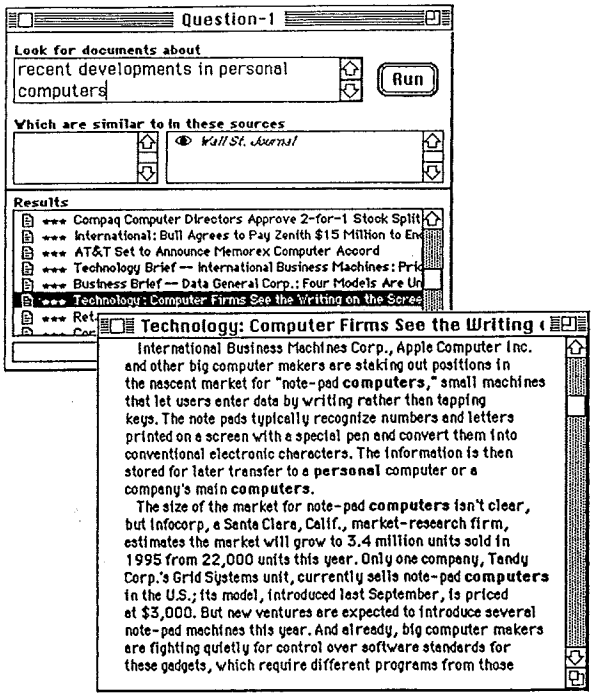
→



Step 1: *Sources are dragged with the mouse into the Question Window. A question can contain multiple sources. When the question is run, it asks for information from each included source.*



Step 2: *When a query is run, headlines of documents satisfying the query are displayed.*



Step 3: *With the mouse, the user clicks on any result document to retrieve it.*

certain set times. When a source of information is contacted, any questions referencing that source are updated with new documents. The users can then easily browse through the results the next morning.

To make the ideal electronic personal newspaper, a system designer would need certain technologies which are not available today. Most computer screens are too small to allow efficient browsing of large amounts of text. Additionally, current data transmission speeds do not allow fast enough scanning if the text is not resident on the user's machine.

Despite current limitations, the WAIS system employs a number of features that will be found in the personal newspaper of the future:

- Clear displays of which questions have new documents
- Searches performed at night to eliminate communications delays
- Documents stored on disk for future reference
- Tools provided to quickly view stored documents

With these techniques, we have established a foundation of user support and acceptance.

SERVERS OR INFORMATION PROVIDERS

The WAIS system was designed to be used by those who wish to sell information, as well as those who want to buy it. It provides a straightforward mechanism for indexing large amounts of data, making it available, and advertising the availability.

The system is flexible enough to provide for a variety of billing methods. A small database producer might make the information available through a telephone connection. Using a 900 number, the billing would be taken care of by the phone company. A slightly more sophisticated site might have a password and credit card billing system. High-volume servers might want to set up flat fee contracts with customers. Other methods will certainly emerge as use increases. The system was designed to be as adaptable as possible to future financial arrangements.

As the dissemination of information becomes easier, questions of ownership, copyright, and theft of data must

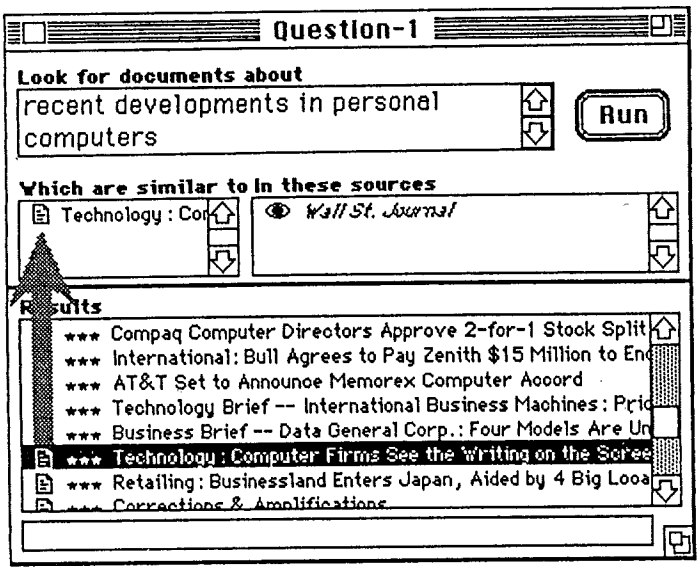
be addressed. These issues confront the entire information processing field, and are particularly acute here. The WAIS system is designed to keep control of the data in the hands of the servers. A server can choose to whom and when the data should be given. Documents are distributed with an explicit copyright disposition in their internal format. This is not to say that theft can not occur, but if a client starts to resell another's data, standard copyright laws can be invoked.

THE DIRECTORY OF SERVERS

As the WAIS system develops, sources of information will proliferate, making it impossible for any user to keep track of all servers that may be available at any one time. To help solve this problem, Thinking Machines is maintaining a Directory of Servers in a widely accessible location. The Directory of Servers contains indexed textual descriptions of all known servers. It is queried just like any other source. Instead of text documents, however, it returns source structures, which are specially formatted files that can be plugged into a question and used for queries.

For example, suppose you needed information concerning the current gross national product of Mali, but had no idea where to find it. You might first ask the directory of servers for "information about the current economic condition of Mali." The directory would return several documents, among which might be a source for the World Factbook, an online almanac maintained by the CIA. You would then use this document as the source field of a question, and re-run the query. This time, the system would contact the almanac, ask for the information, and return a document with the data you need.

Additionally, the Directory of Servers provides a means for information providers to advertise the availability of their data. When a new source becomes available, the developers can submit a textual description, along with the necessary information for contacting the server. This information is added to the directory, and becomes available to the public.



Step 4: *To refine the search, any one or more of the result documents can moved to the* ***Which are similar to:*** *box. When the search is run again, the results will be updated to include documents which are "similar" to the ones selected.*

A COMMON PROTOCOL FOR INFORMATION RETRIEVAL

One of the most far reaching aspects of this project is the development of an open protocol. The four companies have jointly specified a standard protocol for information retrieval. Creating a market where new servers can be readily established requires an open, publicly available protocol. Ideally this protocol would be internationally standardized, yet flexible enough to adapt to new ideas and technologies; functioning over any electronic network, from the highest speed optical connections to phone lines.

The use of an open and versatile protocol fosters hardware independence. This not only provides for a much wider base of users, it allows the system to seamlessly evolve over time as hardware technology progresses. It provides an incentive to produce the best components possible. For example, the protocol provides for the transmission of audio and video as well as text, even though at present most workstations are unable to handle them. However, they are free to ignore pictures and sound returned in response to questions, and to display and retrieve only text. This inability, though, does not hinder higher-end platforms from exploiting their greater processing power and network bandwidth.

The WAIS protocol is an extension of the existing Z39.50 standard from NISO [3]. It has been augmented where necessary to incorporate many of the needs of a full-text information retrieval system [4]. To allow future flexibility, the standard does not restrict the query language or the data format of the information to be retrieved. Nonetheless, a query convention has been established for the existing servers and clients. The resulting WAIS Protocol is general enough to be implemented on a variety of communications systems.

The success of a WAIS-like system depends on a critical mass of users and information services. In order to encourage development and use, Thinking Machines is not only publishing a specification for the protocol, but is also making the source code for a WAIS Protocol implementation freely available. While this software is available at no cost, it comes with no support. We hope that it will facilitate others in developing servers and clients.

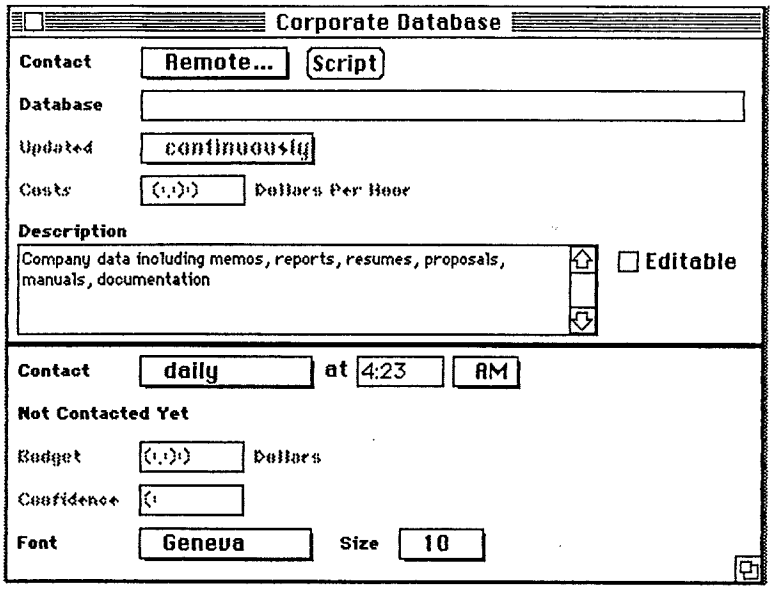
INTO THE FUTURE

In developing the WAIS system, the participating companies have demonstrated that current hardware technology can be effectively used to

→

FIGURE 1
REMOTE INFORMATION SOURCES

The Source description contains all the necessary information for contacting an information server.



provide sophisticated information retrieval services to novice end-users. How this might affect information providers is not yet completely understood. The users at Peat Marwick found the technology useful for day-to-day tasks such as researching potential new accounts and finding resources within their own organization. Since these tasks are not restricted to the accounting and management consulting industries, we are optimistic that this type of technology can be fruitful and productive in many corporate settings.

The future of this system, and others like it, depends upon finding appropriate niches in the electronic publishing domain. Potential uses include making current online services more easily accessible to end-users; or allowing large corporations to access their own internal word processor files more efficiently. It is also possible that near-term development will focus on a single professional field such as patent law or medical research.

REFERENCES

[1] Salton, Gerald and McGill, Micheal. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

[2] DowQuest promotional literature available from Dow Jones & Company, Inc., 200 Liberty Street, New York, NY 10281.

[3] *Z39.50-1988: Information Retrieval Service Definition and Protocol Specification for Library Applications.* National Information Standards Organization (Z39), P.O. Box 1056, Bethesda, MD 20817; 301/975-2814. Available from Document Center, Belmont, CA; 415/591-7600.

[4] Franklin Davis, *et al. WAIS Interface Protocol Prototype Functional Specification*, Thinking Machines. Available from Franklin Davis (Internet — fad@think.com) or Brewster Kahle (Internet — brewster@think.com).

...the protocol provides for the transmission of audio and video as well as text, even though at present most workstations are unable to handle them.

ACKNOWLEDGEMENTS

The design and development of the WAIS Project has been a collective effort, with contributions and ideas coming from many people. Among them are:

Apple Computer: Charlie Bedard, David Casseras, Steve Cisler, Tom Erickson, Ruth Ridder, Eric Roth, John Thompson-Rohrlich, Kevin Tiene, Gitta Soloman, Oliver Steele, Janet Vratny-Watts.

Dow Jones News/Retrieval: Clare Hart, Rod Wang, Roland Laird.

Thinking Machines: Dan Aronson, Franklin Davis, Jonathan Goldman, Chris Madsen, Harry Morris, Patrick Bray, Danny Hillis, Gary Rancourt, Tracy Shen, Craig Stanfill, Steve Swartz, Ephraim Vishniac, David Waltz.

KPMG Peat Marwick: Chris Arbogast, Mark Malone, Tom McDonough, Robin Palmer.

Scolex Information Systems: Art Medlar.

Thanks also to Advanced Software Concepts for TCPack software.

THE AUTHOR

BREWSTER KAHLE has been with Thinking Machines Corporation since it was founded in 1983. He architected the CPU of the Connection Machine Model 2, and led the design of all the custom chips. For the last two years he has been working on making the supercomputer a smart information server in a joint project with Apple, Dow Jones, and Peat Marwick.

Communications to the author should be addressed to Brewster Kahle, Thinking Machines Corporation, 245 First Street, Cambridge, MA 02142; 617-234-1000; or Thinking Machines Corporation, 1010 El Camino Real, Suite 310, Menlo Park, CA 94025; 415-329-9300, Ext. 228; Internet — brewster@think.com.

A REPRINT FROM *ONLINE*®

Online, Inc.
11 Tannery Lane
Weston, CT 06883
203/227-8466